TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Infotehnoloogia eriala

Karol Toompalu

Tõlkekvaliteedi hindamine

Magistritöö (30 EAP)

Juhendaja: Mark Fišel, PhD

Autor: ……..… "….." mai 2011
Juhendaja: ……..… "….." mai 2011

Lubada kaitsmisele
……. "….." mai 2011

TARTU 2011

Sisukord

Sissejuhatus

Masintõlke väljundi hindamine ja veaanalüüs on väga olulised, aga samas raske ülesanne. Käsitsi hindamine on kallis ja aeganõudev. Seetõttu on aastate jooksul uuritud palju erinevaid automaatse hindamise meetrikaid. Kõige laialdasemalt kasutusel olevad meetrikad on sõna veamäär -WER (Word Error Rate), positsioonist sõltumatu sõna veamäär -PER (Position independent word Error Rate), BLEU, NIST ning METEOR. Need meetrikad on osutunud väärtuslikeks tööriistadeks erinevate süsteemide omavaheliseks ning nende süsteemide endi paranduste hindamiseks. Aga samas ei anna need meetrikad täpsemat infot tõlkimise vigadest. Seetõttu tuleb süsteemi väljundile teha detailsemat analüüsi, et tuvastada probleeme süstemis ning suunata uurimistööd.

Magistritöö eesmärk on disainida automaatne tõlkeväljundi veatüüpide analüüs mis mitte ei annaks ühte hinnangut tõlkesüsteemile, vaid esitaks arendajale olulised faktid tõlkesüsteemi võimalike vigade esinemise (sageduse) kohta, mille alusel oleks võimalik teha parandavaid muudatusi tõlkesüsteemi arenduses. Töö esimeses osas tutvustatakse masintõlkesüsteeme ja nendega kaasnevaid probleeme ning mõningaid meetrikaid nende töö hindamiseks. Teises peatükis tutvustatakse lähemalt tõlkimisel tehtavaid vigu ning nende klassifikatsiooni. Viimases peatükis tutvustatakse autori poolt loodud automaatset tõlkekvaliteedi hindajat ning sellega sooritatud katseid.

1. Masintõlge

Masintõlge on arvutilingvistika haru mis uurib arvutitarkvara kasutamise võimalusi teksti või kõne ühest loomulikust keelest teise tõlkimisel.

Kõige lihtsamat tüüpi masintõlke puhul asendatakse ühe keele sõnad teise keele sõnadega sõnastiku abil, arvestamata mingeid keele eripärasid. Saadud tõlge võib olla isegi seda keelt rääkivale inimesele arusaamatu. Arenenumad süsteemid kasutavad tõlkimisel paralleelkorpuseid, reegleid või nende kombinatsiooni. Masintõlke tulemused on pidevalt paranenud, kuid siiamaani pole veel suudetud arendada süsteemi, mis annaks kvaliteetse tõlke ilma sisendteksti piiranguteta. Seega pole arutelud masintõlke edusammude ja potentsiaali üle kaugeltki lõppenud. Tihti on arutletud, kas masintõlke õnnestumiseks tuleks kõigepealt lahendada loomuliku keele mõistmise probleem. Kriitikute arvates tekivad põhimõttelised raskused täisautomaatse masintõlke saavutamiseks [6].

Õigetel tingimustel on masintõlge väga võimas tööriist, sest masintõlge annab küll madala kvaliteediga tõlke, kuid annab selle sekundite või minutitega. Paljudel juhtudel on kasulikum omada sellist inimesele arusaadavat tõlget minutitega, kui kvaliteetset tõlget näiteks nädala möödudes. Selle peatüki materjal põhineb allikatel [4], [7], [10], [11] ja [16].

1.1. Lähenemisviisid

Masintõlke probleemi lahendamiseks on välja mõeldud mitmeid lähenemisviise. Ajalooliselt on vanim vahekeele ehk interlingua abil tõlkimine mis pakuti välja juba 17-ndal sajandil. Sellest hoolimata hakati üha enam uurima mingi kahe konkreetse keele omavahelist tõlkimist. Esimesed süsteemid, mis töötasid arvutitel olid reeglipõhised. Viimasel ajal on üha rohkem kasutusel statistilistel meetoditel põhinevad süsteemid.

1.1.1. Reeglitepõhine lähenemine

Reeglitepõhine lähenemine, samuti tuntud ka kui teadmistepõhine lähenemine põhineb lingvistilisel informatsioonil, mis saadakse (kakskeelsetest) sõnastikest ja grammatikatest, hõlmates põhilisi semantilisi, morfoloogilisi ja süntaktilisi reeglipärasusi. Tõlge genereeritakse lähtekeele ja sihtkeele semantilise, morfoloogilise ja süntaktilise analüüsi alusel.

Interlinguaalne meetod

Interlinguaalse meetodi puhul sünteesitakse lähtetekst algul mingile vahekujule, mis peaks olema lähtekeele teksti mõtte loomulikest keeltest sõltumatu nö interlinguaalne esitus, ning

selle esituse järgi konstrueeritakse vastav sihtkeelne tekst. Interlinguaalne esitus peab olema selline, et sihtteksti on võimalik selle järgi konstrueerida ilma algtekstile ega lähtekeelele vaatamata. Interlinguaalne esitus on nii sisend- kui väljundteksti ühine abstraktne esitus.

Interlinguaalse meetodi kasutamisel on siiski olulisi puudusi. Interlinguat on väga raske defineerida ja seda isegi lähedaste sugulaskeelte vaheliseks tõlkimiseks. Päris universaalset interlinguat, mille abil oleks võimalik tõlkida suvalisest keelses suvalisse keelde, pole veel suudetud leida.

Ülekandemeetod

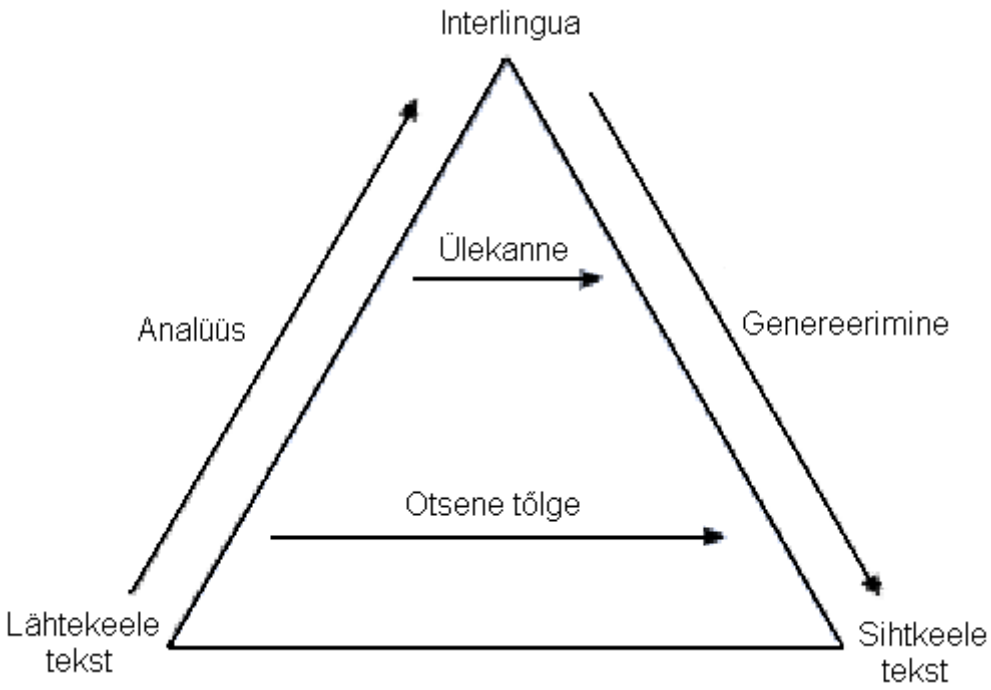
Ülekandemeetodi puhul kasutatakse ühest keelest teise keelde tõlkimisel kahte vaheesitust. Esimene saadakse algteksti analüüsil ning teise järgi konstrueeritakse sihtkeelne väljund. Ühest vaheesitusest teise teisendamiseks on iga keelepaari jaoks loodud nn ülekandeblokk. Vaheesitused on tugevasti keelest sõltuvad - esimene lähtekeelest, teine sihtkeelest. Esimene vaheesitus on lähtekeelse, teine sihtkeelse teksti abstraktne esitus. Keelest sõltumatuid esitusi ülekandemeetodil ei genereerita.

Ülekandemeetodi tasemed on pealiskaudne (süntaktiline) ja sügav (semantiline). Pealiskaudne ülekandemeetod kannab üle süntaktilist struktuuri, sobilik sama perekonda kuuluvate keelte tõlkimiseks. Sügav ülekandemeetod konstrueerib semantilise esituse mis sõltub sihtkeelest. Esitus koosneb seeriast tähendust esitavast struktuurist. Tavaliselt moodustab süsteem predikaadid. Enamasti on ka vajalik ka struktuurilist ülekandmist.

Sõnastikupõhine meetod

Sõnastikupõhise meetod on kõige lihtsakoelisem, selle korral tõlgitakse tekst sõnastiku sissekannete alusel, sõna sõnalt. Sobib paremini lühemate fraaside kui tervete lausete tõlkimiseks. Meetod pole eriti täpne, kuid ka sellel on omad väljundid. Näiteks on väljundtekst abiks käsitsi tõlkimisel. Lisaks sõnastiku kasutamisele võib defineerida reeglid, mis muudavad järjekorda, lisavad näiteks funktsionaalseid sõnu. Need reeglid võimaldavad meetodil olla kasulik ka laiemate valdkondade tekste tõlkimisel.

Joonis 1. näitab lähtekeele analüüsi ühes küljes ja sünteesi sihtkeelde teises küljes. Püramiidi tipp näitab teoreetilist kontekstivaba väljendust, mis on saavutatud teksti analüüsiga ja on sobilik otsesünteesiks. Teekond püramiidi tippu on pikk. Diagramm peaks näitama ka seda, et mida rohkem on teksti analüüsitud, seda vähem on vaja vaeva näha ülekandmisel. Ekstremaalne olukord on püramiidi põhi, kus kogu töö tuleb ära teha ülekandega, sest analüüsi pole peaaegu üldse teostatud.



Joonis 1. Vaheesituste sügavusi kujutav püramiid.

1.1.2. Statistiline meetod

Statistilise meetodi[1] korral genereeritakse tõlge statistiliste mudelite abil, mis on tuletatud kakskeelse korpuse analüüsist. Tekst tõlgitakse tõenäosusjaotuse järgi, et antud sõne lähtekeeles on tõlge sihtkeeles.

Statistilisel masintõlkel on mitu eelist traditsiooniliste suundade ees. Näiteks annab ta enamasti loomutruuma tõlke ning kasutab paremini ära olemasolevaid keeleressursse, mille jaoks pole vaja kasutada kallist inimressurssi, samas on aga statistilise mudeli treenimiseks vaja koguda ja märgendada korpuseid, mis on töömahukas ja veaohtlik.

1.1.3. Näitepõhine meetod

Näitepõhise meetodi[12] põhiidee on tõlkimine analoogia alusel: kui kord juba tõlgitud lause esineb tekstis taas, siis on tõenäoline, et selle võib tõlkida samuti kui enne.

Tõlkimisel ei kasutata täpseid vastavusreegleid, vaid lause tõlkimiseks valitakse kakskeelses andmebaasis olemasolevate lähtekeelsete lausete hulgast lähtelause fraasidele kõige lähedasemad tõlked. Spetsiaalse mehhanismi abil, mis kasutab statistilisi sõltuvusi ja heuristilisi reegleid, kombineeritakse need fraasid lauseks.

1.1.4. Hübriidsed meetodid

Hübriidse meetodi korral kasutatakse ära statistilise ja reeglipõhise meetodi tugevused. Lähenemise viise on erinevaid:

- **Reeglid järeltöödeldakse statistikaga**: Tõlge sooritatakse reeglipõhise mootori abil, seejärel kasutatakse statistikat, et korrigeerida selle väljundit.
- **Reeglite poolt juhitud statistika**: Reegleid kasutatakse andmete eeltöötluseks, et parandada statistikamootorit. Reegleid kasutatakse ka statistilise meetodi väljundi järeltöötluseks, et sooritada ülesandeid nagu näiteks normaliseerimine. Sellel rakendusel on tõlkimise ajal palju rohkem jõudu, painduvust ja kontrolli.

1.2. Masintõlkega esile kerkivad probleemid

Paljud loomuliku keele tõlkimise automatiseerimisel tekkivad probleemid on sarnased loomulike keelte mõistmise probleemidega. Selleks, et keelt tõlkida, tuleb sellest aru saada. On olemas palju vahendeid erinevate keelte analüüsiks ja sünteesiks, kuid siiski pole veel ühtegi süsteemi, mis suudeks seda teha sama hästi, kui inimene. Lisaks tekivad ka probleemid juhul, kui tõlgitavad keeled erinevad teineteisest väga palju lause ülesehituse või teiste keele eripärade poolest.

Ühestamine

Sõnade kahemõttelisus põhjustab probleeme nii lähtekeele kui sihtkeele analüüsi käigus. Näiteks kui lähtekeele sõna on kahemõtteline tekib probleem, milline sõna tähendus valida tõlkimiseks.

Selle probleemi lahendamiseks on palju lähenemisviise, need saab orienteeruvalt jagada pealiskaudseks ja sügavaks lähenemisviisiks. Pealiskaudne lähenemisviisis ei võta tekstist arvesse mingit teadmust. Selle korral lihtsalt kasutatakse statistilist meetodit sõna ümbrusele. Sügavad lähenemisviisid eeldavad põhjalikke teadmisi sõnadest. Praktikas pole sügav lähenemisviis eriti edukas, sest selliseid teadmisi pole veel arvutile arusaadavas formaadis. Samas kui need oleksid olemas, oleks sügav lähenemisviis palju täpsem, kui pealiskaudne lähenemisviis.

Olemite tuvastamine

Tekstis atomaarsete elementide tuvastamine eeldefineeritud kategooriatesse, näiteks pärisnimed, kogused, raha suurus, protsendid jne.

Probleemid tekivad, kui tõlkesüsteem kohtab tundmatut sõna/sõnapaari, mis on näiteks pärisnimi ning ajab selle sassi sageli esineva loomuliku keele sõnaga.

Morfoloogiliselt rikkad keeled

Statistilisele masintõlkele pakub suurt väljakutset tõlkimine keelte vahel, mille morfoloogilised struktuurid erinevad teineteisest suurel määral. Kui tõlkida morfoloogiliselt rikkast keelest tekivad probleemid sihtkeelele tundmatud ning võivad olla morfoloogiliselt mitmesed. Teist pidi tõlkimisel tekib raskusi õige sõnavormi valimisel ja genereerimisel.

Keelte süntaktilised/grammatilised erinevused

Erinevused keelte sõnade järjestuses ja lausete ehituses põhjustavad vale sõnade järjekorra ja lause ehituse vigade tekke sihtkeeles.

1.3 Masintõlke süsteemide hindamine

Masintõlkesüsteemide hindamisel lähtutakse järgnevatest kriteeriumitest:

- Puhta teksti tõlke kvaliteet: arusaadavus ja mõistetavus, täpsus, usaldatavus, stiil;
- Rakendatavus sõnastike koostamiseks ning täiendamiseks, järelkontrolliga tekstidele, jms.;
- Laiendatavus uute keelepaaride ning uute ainevaldkondade jaoks;
- Kulutused võrreldes inimese poolt läbiviidava tõlkega.

Automaatsete hindamismeetodite testimisel kasutatakse vastava ainevaldkonna tüüpilisi tekste, mille tõlkeid võrreldakse inimese poolt tõlgitutega.

Automaatse hindamise meetodid on märgatavalt kiirendanud masintõlke süsteemide arendustsüklit. Nad etendavad võtmerolle, lubades kiiret numbrina esitatavat tõlke hinnangut, mis abistab süsteemide arendajaid nende igapäevaste otsuste langetamisel. Kuid praeguste hindamismeetoditel on ka omad puudused, näiteks ei suuda nad esitada usaldusväärset hinnangut lausete tasemel. Lisaks ei esita nad mingit täpsustavat infot leitud vigade kohta, mida läheks vaja arendajatel oma süsteemi tugevuste ja vigade leidmisel.

Masintõlke arenduse kontekstis on hindamismeetodite ülesanneteks:

- Veaanalüüs – leida ja analüüsida võimalikke veatüüpe. Täpsed teadmises süsteemi võimetest on vajalikud selle käitumise parandamiseks.
- Süsteemide võrdlus – mõõta välja pakutud muudatuste efektiivsust võrreldes sama süsteemi erinevaid versioone.
- Süsteemi optimeerimine – algparameetrite kohandamine, et maksimeerida süsteemi üldist kvaliteeti.

1.4. Tõlkekvaliteedi meetrikad

Masintõlke puhul on mitmeid erinevaid ja võrdselt täpseid viise lause tõlkimiseks. Samuti saab osalauseid lauses ümber tõsta. Lahenduseks oleks paljude testlausete võrdlemine, mitme tõlke lisamine ühe lause ideaaltõlkeks, fraaside/ n-grammide otsimine lubades nende ümberpaigutust.

Kõige levinumad automaatsed tõlkekvaliteedi meetrikad on BLEU, NIST ja METEOR,

1.4.1. WER ja PER

WER (*word error rate*) – sõna veamäär põhineb Levenshtein kugusel[8] – minimaalne arv sõnade asendamisi, kustutamisi ja lisamisi, mis tuleb teha, et saada lähtelausest sihtkeeles olev lause. Sõna veamäära puudujääk esineb selles, et ta ei luba sõnade ümberjärjestamist, sest tõlge võib olla korrektne ka isegi siis, kui sõnade järjekord erineb tõlkenäitest. Tõlge pole lineaarne, üksüheselt vastav inimtõlkele.

WER = (asendamised + lisamised + kustutamised)/ soovitusliku tõlke pikkus

WER põhiprobleemi lahendamiseks on välja töötatud positsioonist sõltumatu sõna veamäär [PER (*posotion independent* WER)], mis võrdleb lauses esinevaid sõnu sõnade järjekorrale tähelepanu pööramata. PER on alati väiksem või võrdne WER väärtusega. Samas on ka PER meertikal oma puudus – sõnade järjekord võib olla mingis kontekstis vägagi oluline. Seega parim lahendus oleks arvutada mõlemad veamäärad.

Kui WER ja PER skoorid on teineteisest väga erinevad, siis on tõlkes palju sõnade järjestusvigu. Kui PER skoor ja PER skoor sama lause sõnade algvormidest on väga erinev, siis on tõlkes palju sõnalõppude vigu.

1.4.2. BLEU

Hetkel kõige populaarsem meetrika on BLEU[5], mis pakub lahenduse sõnade järjekorra probleemile. See töötab sarnaselt PER meetrikale, aga võtab arvesse pikemate n-grammide kattuvust soovitusliku tõlkega. Andes n-gramm vasted, saame leida n-gramm täpsuse – kindlas astmes olevate n-grammide suhe kõikide samas astmes genereeritud n-grammidega.

n-gramm on võrdlemise mõõtühik, ehk unigramm puhul arvestatakse üksikuid sõnu, 2-gram puhul arvestatakse kahte järjestikust sõna jne.

Täpsusel põhinevate meetrikate probleemile – puuduvate sõnade mitte hindamine, on BLEU vastanud lühiduse hinna lisamisega.

$$BLEU = \text{Lühiduse hind} \cdot \exp\left(\sum_n \lambda_i \cdot \log\left(\text{täpsus}_n\right)\right)$$

$$\text{Lühiduse hind} = \min\left(1, \frac{\text{väljundi pikkus}}{\text{sisendi pikkus}}\right)$$

$$\text{Täpsus} = \frac{\text{õiged sõnad}}{\text{väljundi pikkus}}$$

Tavaliselt on maksimaalne n-grammide aste 4. Seda nimetatakse BLEU-4. Kusjuures on tavaliselt kaaluks λ_i seatud 1.

BLEU eeliseks on hea korreleeruvus inimeste hinnangutega, kuid puuduseks on see, et pikemad n-grammid domineerivad lühemaid n-gramme, näiteks kui 2-gramme on antud tõlkes palju vähem, kui unigramme, siis saab süsteem väga kehva tulemuse.

BLEU on üks esimesi meetrikaid, mis on näidanud head korrelatsiooni inimeste kvaliteedi hinnangutega.

1.4.3. NIST

NIST meertika[9] põhineb BLEU meetrikale, aga väikeste erinevustega. Kui BLEU kalkuleerib n-gramm täpsuse kus kõik on võrdse kaaluga. Siis NIST arvutab lisaks ka iga n-grammi informatiivsuse, mida haruldasem korrektne n-gramm leitakse, seda suurem kaal talle omistatakse ning erinevalt BLEU meetrikast arvutab NIST aritmeetilise keskmise. NIST erineb ka lühiduse hinna arvutamise poolest, kuid see ei muuda oluliselt lõppskoori.

1.4.4. METEOR

METEOR meetrika[14] on disainitud mõningaid BLEU puudusi eemaldama. Meetrika põhineb unigrammi täpsuse ja saagikuse kaalutud harmoonilisel keskmisel. Meetodile on lisatud ka funktsioone, mida pole teistes meetrikates, näiteks sünonüümide vastavusse seadmine. Lisaks sisaldab ka sõnatüvede leidjat, mis leiab sõnadele lemmad ning viib vastavusse sõnavormide lemmad.

1.5. Meta-hindamine

Automaatsed hindamismeetodid võimaldavad teadlastel anda hinnangut ja optimeerida oma süsteeme ilma kallist inimressurssi kasutamata. Samas tekitab automaatsete hindamismeetodite kasutamine arendamisse lisa sammu – üldhindamine ehk hindamismeetodi hindamine, sest arenduseks valitud hindamismeetod juhendab arendajat, leides puudusi süsteemis ning selle abil hindab arendaja, kas tehtud muudatused olid kasulikud või mitte.

Hindamismeetodite hindamise eesmärk on uurida, kui hästi korreleerub hindamismeetod inimhinnangutega süsteemi tasandil ning lause tasandil.

Enamasti hinnatakse hindamisemeetodeid adekvaatsuse või sõnaosavuse hinnangute vastu või kombinatsiooniga neist, kasutades kas Pearson, Spearman või Kendall korrelatsiooni-kordajaid:

> Pearson correlation coefficent[15] mõõdab kahe muutuja vahelise lineaarse seose tugevust.
>
> Measure Spearman's rank correlation[15] hindab kui hästi kahe muutuja vahelist seost on võimalik väljendada monotoonse funktsiooniga.
>
> Kendall rank correlation coefficient[15] hindab kahe muutuja vahelise sõltuvuse tugevust.

2. Tõlke vigade analüüs

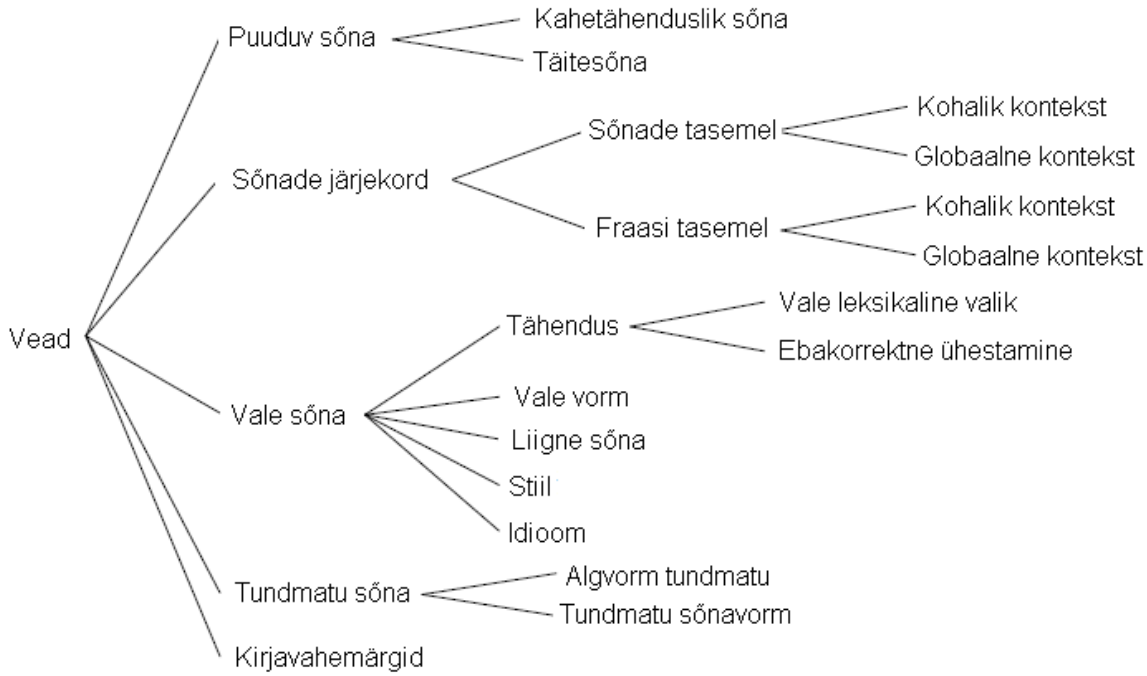
Selleks, et leida vigu tõlkes on vaja ühte või mitut soovituslikku tõlget, mida saaks võrrelda masintõlke väljundiga. On teada, et igal lausel on mitu õiget tõlget, mis osutub raskeks probleemiks automaatsel hindamisel ja süsteemide võrdlusel.

Sageli põhjustab üht tüüpi vea esinemine teist tüüpi vea tekke. Näiteks kehv sõna tõlge võib põhjustada kehva sõnade järjestuse genereeritud lauses.

2006. aastal pakuti artiklis „*Error Analysis of Statistical Machine Translation Output*" David Vilar jt. poolt välja järgnev vigade klassifikatsioon, mis on hetkel ainus laialt tuntud klassifikatsioon [2].

2.1. Vigade klassifikatsioon

Masintõlke vigade klassifikatsioon on lihtsa hierarhilise struktuuriga, mis on näidatud Joonise 2. Esimesel tasemel on puuduvad sõnad, valed sõnad, tundmatud sõnad, sõnade järjekord ja kirjavahemärgid.



Joonis 2. Vigade klassifikatsioon.

2.1.1. Sõnade järjekord

See kategooria pühendub genereeritud lause sõnade järjekorrale. Eristada saab sõna ja fraasi tasemel ümberjärjestust. Sõnade tasemel ümberjärjestamisel saab korrektse lause

liigutades üksikuid üksteisest sõltumatuid sõnu, fraasi tasemel ümberjärjestamisel liigutatakse kõrvuti asetsevate sõnade blokke. Lokaalse ja globaalse konteksti suurust on raske kindlalt määrata, aga selle abil püütakse väljendada sõnade liigutamist lokaalses kontekstis (süntaktilises tükis) või erinevate süntaktiliste tükkide vahel.

Sõnade järjekorda on võimalik automaatselt kontrollida, kuid komistuskiviks võib osutuda kombinatsioon teiste vigadega, näitaks sõna morfoloogilise info valesti tõlkimine.

2.1.2. Puuduvad sõnad

Puuduv sõna on viga, mis tekib, kui genereeritud lauses puudub sõna, mis esineb soovituslikus tõlkes. Puuduv sõna võib olla vajalik lause tähenduse väljendamisel, või võib olla nn. täiend sõna, mis väljendab lause grammatilist vormi kuid lause tähendus on säilitatud. Esimest tüüp sõnade vead on olulisemad ning tuleks esimesena lahendada.

Puuduvaid sõnu on lihtne leida, kuid sõnade olulikkust lause tähenduses ei ole võimalik määrata ilma morfoloogilise ja semantilise infota.

2.1.3 Vigased sõnad

Kõige laiem kategooria on vigased sõnad. Need tekivad, kui süsteem ei suuda antud sõnale leida korrektset tõlget. Vigastel sõnadel on mitu alamkategooriat. Esimese puhul rikub ebakorrektne sõna lause tähendust, mille puhul tekkis viga kas sõnale tõlke valimisel või lähtekeeles sõnale korrektse tähenduse valimisel. Teine alamkategooria vead tekivad sõnavormi valikul, kuigi sõna tüvi oli õige. Järgmise kategooria vead tekivad liigsete sõnade genereerimisel tõlkesse. Viimased kaks kategooriat on vähetähtsad. Stiili vigade puhul on lause tähendus edasi antud, aga sõnadevalik on kehv. Näiteks on sõnade kordamine stiili viga. Viimaks on väljendid, mida süsteem ei tunne (idioomid) ning mis tõlgitakse tavalise tekstina. Tavaliselt ei saa neid otse tõlkida, mis põhjustab tõlkes veel vigu.

2.1.4 Tundmatud sõnad

Tundmatud sõnad on samuti vigade allikaks. Siin saab eristada sõnu, mille korral on selle tüvi (või lemma) süsteemile tuntud, aga sõnavorm tundmatutu ning täiesti tundmatuid sõnu.

2.1.5. Kirjavahemärgid

Viimasena saab eristada kirjavahemärkide vigu, aga praeguse masintõlke kvaliteedi puhul põhjustavad need ainult väikseid häireid ilma fikseeritud kirjavahemärgi reegliteta keeltes ning neid ei uurita rohkem selles töös.

3. Automaatne tõlkevigade hindaja

Automaatse tõlkevigade hindaja eesmärgiks on võrrelda masintõlke ja inimtõlke väljundeid ning esitada detailselt erinevused. Programm pöörab tähelepanu sõnade/fraaside järjekorrale, puuduvatele, liigsetele ja ebakorrektsete sõnade/fraaside esinemise hulgale Selleks leitakse iga lause kohta eraldi vastav statistika ning lõpptulemusena esitatakse lausete statistikate aritmeetilised keskmised. Eesmärk on anda arendajale võimalikult palju infot tema süsteemi käitumise kohta, et arendaja saaks ise anda hinnangu süsteemi poolt tekitavate vigade kohta.

3.1. Ülesande lahenduseks loodud vahendid

Käesolevas töös püstitatud ülesande lahendamiseks loodi mitme mooduliga Python programm. Valitud sai modulaarne lähenemissuund, kus moodulid on enamasti lühikesed ja mingi konkreetse ülesande lahendamiseks, et lihtsustada programmist arusaamist ning koodi taaskasutamist. Programmi põhikeeleks sai Java asemel valitud Python, sest sellel puudub kompileerimisvajadus ning mudatuste tegemine on seega lihtne ja kiire. Lisaks saab Python listidesse panna ebaregulaarseid andmeid, mis lihtsustab oluliselt loomuliku keele töötlust.

Loodud programm ja selle moodulid on kirjeldatud peatüki lisa 1 all ning kaasatud ka tööle lisatud CD peal olevas arhiivis.

3.2. Programmi töö etapid

Programmi töö on jaotatud etappidesse, kus iga etapi puhul leitakse hinnang konkreetset tüüpi tõlkimisvigade esinemisele.

Programmi väljundi selgitus

Programm raporteerib sõnade veamäära ja positsioonist sõltumatu veamäära, mille puhul on esitatud protsent tõlgete kattuvus, ehk mida suurem on raporteeritud arv, seda parem on tõlge. Sõnade veamäära ja positsioonist sõltumatu veamäära abil saab näiteks hinnata sõnade järjekorra korrektsust: mida suurem on nende meetrikate erinevus, seda rohkem esineb tõlkes sõnade järjestuse vigu.

BLEU meetrika hinnang korreleerub hästi inimhinnangutega ja seega saab selle abil ligikaudu sarnase hinnangu inimese hinnangule.

Liigsed sõnad näitavad, kui suur protsent masintõlkest on liigsed. Puuduvad sõnad näitavad, kui suur protsent inimtõlke sõnadest ei ole esindatud masintõlkes. Neid kahte

arvu suurendab teine tõlkevea tüüp: sõna vale tõlge. Täpselt kattuvad protsent näitab, kui suur osa inimtõlke sõnadest on esindatud masintõlkes.

Keskmine puuduvate/liigsete n-grammide arv lausetes kirjeldab, kui palju n-gramme pikkusega 1 kuni 4 on igas lauses keskmiselt puudu/liigselt.

N-grammide järjekord kirjeldab lausete sõnade järjekorra olukorda. Näiteks unigrammide puhul on vaatluse all üksikud sõnad ja bigrammide puhul sõnade paarid.

Viimasena esitab programm sarnaste sõnatüvedega esinevate sõnade hulga tõlkes, mille abil saab määrata sõnalõppude tõlkevigade sagedust. Programmile on ette antud veamäär 25%, mis tähendab, et 25% sõna lõpust võib erineda inimtõlkest. Näiteks „use" ja „uses" ühisosa on „use" ning muuta on vaja ühte tähte neljast, seega viga on ¼ ehk 25%.

3.2.1. Eeltöötlus

Tekste töödeldi eelnevalt, et eemaldada kirjavahemärgid. Sisend on struktureeritud nii, et iga lause on erineval real ning võrreldavates tekstides on ühel real samale sisendlausele vastavad tõlked.

3.2.2. Puuduvad ja liigsed sõnad

Puuduvad sõnad on sõnad, mis esinevad inimtõlkes, kuid ei esine masintõlkes. Liigsete sõnadega on vastupidi - nad esinevad masintõlkes, kuid puuduvad inimtõlkes. Seda tüüpi sõnade leidmine ei tohiks olla raske: võtta ühe lause/teksti sõnad ja eemaldada sealt sõnad, mis esinevad teises lause/tekstis. Selle lähenemise korral tekib probleem: kirjeldamata jäävad sõnad, mis esinevad küll masintõlkes ja inimtõlkes, aga nende esinemise arv on erinev. Näiteks kui sõna esineb inimtõlkes 2 korda ning masintõlkes 1 kord:

Masintõlge: *It was the best times.*

Inimtõlge: *It was the best of times, it was the worst of times.*

Selle näite puhul on ainukesed puuduvad sõnad „*of*" ja „*worst*", samas kui lause mõte on hoopis erinev. Samas on ka teiste sõnade esinemise arv erinev, mida tuleks kindlasti arvesse võtta süsteemi hindamisel.

Selle probleemi lahenduseks on programmis leitud lisaks üks kord esinevate puuduvate ja liigsete sõnade hulgale ka mitu korda esinevate puhul sõnade hulgad, mis esinevad inimtõlkes ja masintõlkes erinev arv kordi ning need puuduvate/liigsete vastavad hulgad summeeritud. Näiteks eelneva näite korral esineb sõna „*times*" kaks korda inimtõlkes ja üks kord masintõlkes, ning kajastatakse statistikas ühe puuduva sõnana.

Fraaside puhul käitub kõik sama moodi, kui üksikute sõnadega, arvestusse on lihtsalt võetud mitu sõna järjestikku.

3.2.3. Sõnade/fraaside järjekord

See probleem on kõige laiahaardelisem ja raskem, sest lisaks sõnade järjekorrale võivad tõlkes esineda ka puuduvad/liigsed sõnad ning sõnade tõlkimisel tekkivad vead, mis ajavad sõnade järjekorra veel rohkem sassi ning täpse hinnangu andmise raskemaks.

Sõnade/fraaside järjekorra hindamiseks on programmis leitud hinnangud erineva pikkusega n-grammidele. Nende abil saab hinnata sõnade-fraaside tasemel esinevate järjekorra vigu.

Näide:

Tõlge 1: *The cat sat on the mat.*

Tõlge 2: *On the mat the cat sat.*

Selle näite puhul on mõlemad fraasid „*on the mat*“ ja „*the cat sat*“ keeleliselt korrektsed, kuid saadud laused teineteisest erinevad.

3.2.4. Vale sõna tõlge

Tõlkimisel vale sõna valiku tuvastamine on raske, sest pole teada, milline sõna lähtelauses asendati sihtkeeles oleva sõnaga. Seega on pea ilmvõimatu panna inimtõlke ja masintõlke sõnu vastavusse ilma konkreetse keele ressursse, näiteks sõnastike või sünonüümisõnastikke kasutamata. Nende ressursside kasutamine on töös limiteeritud, et vähendada keelest sõltuvust.

Kõige lihtsam on leida tõlkevigu, kus sõna tüvi on õigesti tõlgitud, aga sõna lõpp valesti. Selle vea avastamiseks on kasutusele võetud tõlketekstide morfoloogilised analüüsid. Samuti on proovitud kasutada sõnatüvede keelest sõltumatut tuvastamist lihtsa sõna veamäära abil.

Järgneva näite korral on tõlgitud sõna „*cat*“ õigesti, kuid valesti on sõna lõpp. Ilma morfoloogiliselt analüüsitud teksti või veamäära arvestamata, klassifitseeritaks sõna „*cats*“ liigseks sõnaks ning „*cat*“ puuduvaks sõnaks:

Masintõlge: *The cats sat on the mat.*

Inimtõlge: *The cat sat on the mat.*

3.3. Programmi realiseerimise käigus tekkinud probleemid

Programmi arendamise käigus sai üha selgemaks, et täielikku veaanalüüsi on võimatu automaatsete meetoditega lahendada. Lisaks keelespetsiifilisema probleemi lahendamiseks tuleb tuua sisse üha rohkem keeleressursse, millest proovitakse selle töö raames võimalikult palju hoiduda. Samas oli võimalik näiteks sõnade vale vormi tõlke tuvastamine suhteliselt kerge sisse tuua, sest on olemas üpris hästi töötavad morfoloogilised analüsaatorid. Samas kui idioomide ja stiilivigade puhul tuleks sisse tuua palju erinevaid keeleressursse ning nende tuvastamine isegi siis raske. Sellest tulenevalt ei tuvasta programm sõna ühestamisel ja leksikaalsel valikul tekkinud vigu ning süsteemile tundmatuid sõnu, need klassifitseeritakse vale/liigse sõna klassi.

3.4. Katsed

Eestis keskendutakse enamasti inglise-eesti-inglise tõlkimisele, seega tehti ka katsed kahe keele, eesti ja inglise keele peal.

Antud katsete eesmärk on demonstreerida programmi kasutatavust ja rakendusvõimalusi.

Katsete puhul hinnati sama teksti tõlget mitme erineva masintõlke vastu.

Lisaks eelmainitud tüüpvigade leidmisele arvutati ka tõlke korreleerumise hindamiseks kolme meetrika hinnangud: BLEU, sõna veamäär, positsioonist sõltumatu sõna veamäär.

Kõige tavalisem on arvestada kuni 4-gramm täpsusega fraase ning sama täpsust kasutati ka programmi testimisel. Järgnevate katsete puhul on sisendist eemaldatud kirjavahemärgid, kõik muu on originaalkujul.

3.4.1. Eesti keelde tõlgitud teksti hindamine

Inglise keelest eesti keelde tõlkimisel on sisendina kasutatud väiksest osa JRC-Acquis v.3.0[13] korpusest.

Tõlkimise vahendina on kasutatud paketti Joshua[17] kahte versiooni.

Tõlke sisend koosneb 2500 lausest, 81468 sõnast ning inimtõlge koosneb 58566 sõnast.

Joshua imb2rl

```
Tulemused:

korreleerumine inimtõlkega
Sõna veamäär (kattuvuse %) 37.8736
Positsioonist sõltumatu sõna veamäär (kattuvuse %) 55.2884
BLEU meetrika 0.1681

puuduvad/liigsed unigrammid:
Liigsed 34.4765 %
Puuduvad 35.0207 %
Täpselt kattuvad 58.4331 %
keskmine puuduvate/liigsete n-grammide arv lausetes :
Liigseid:
unigramme 8.8992, bigrammme 21.9532, trigramme 23.4860, 4-gramme 23.9648
Puuduvaid:
unigramme 8.6828, bigrammme 21.5304, trigramme 23.0540, 4-gramme 23.5292

n-grammide järjekord:
Masintõlkes korrektseid n-gramme osakaal keskmiselt:
unigramme 0.6006, bigrammme 0.3656, trigramme 0.2545, 4-gramme 0.1913
Masintõlkes korrektsete n-grammide keskmine osakaal inimtõlkes:
unigramme 0.5957, bigrammme 0.3626, trigramme 0.2523, 4-gramme 0.1890

Vale sõna - sõna tõlkimisel tekkivad vead

Osaliselt kattuvad (sarnaste tüvedega) 6.6057 %
```

Hinnang:

Tõlge korreleerub päris hästi inimtõlkega, kuid probleeme on n-grammide saagisega, näiteks tõlgib süsteem õigesti ainult ligikaudu 19 protsenti fraasidest pikkusega 4 sellest hoolimata, et 60 protsenti tõlkes sisalduvatest sõnadest on korrektsed. Ligi 7 protsenti sõnadest kattuvad osaliselt, ilmselt on tegu vale tüve lõpuga.

Joshua hmm

```
Tulemused:

korreleerumine inimtõlkega
Sõna veamäär (kattuvuse %) 39.8906
Positsioonist sõltumatu sõna veamäär (kattuvuse %) 56.9369
BLEU meetrika 0.1838

puuduvad/liigsed unigrammid:
Liigsed 32.8828 %
Puuduvad 33.7696 %
Täpselt kattuvad 59.7593 %
keskmine puuduvate/liigsete n-grammide arv lausetes :
Liigseid:
unigramme 8.6124, bigrammme 21.3388, trigramme 22.9060, 4-gramme 23.4628
Puuduvaid:
unigramme 8.3200, bigrammme 20.7500, trigramme 22.3176, 4-gramme 22.8732
```

```
n-grammide järjekord:
Masintõlkes korrektseid n-gramme osakaal keskmiselt:
unigramme 0.6160, bigrammme 0.3828, trigramme 0.2725, 4-gramme 0.2064
Masintõlkes korrektsete n-grammide keskmine osakaal inimtõlkes:
unigramme 0.6088, bigrammme 0.3791, trigramme 0.2703, 4-gramme 0.2039

Vale sõna - sõna tõlkimisel tekkivad vead

Osaliselt kattuvad (sarnaste tüvedega) 6.5217 %
```

Hinnang:

Tõlge on natuke parem eelnevast tõlkesüsteemi tõlkest, kuid süsteemil on sarnased vead: väiksemate ühikute täpsus on tunduvalt suurem kui pikemate ühikute puhul. Ka osaliselt kattuvate sõnade arv on sarnane eelmise tõlkemudeli väljundile.

3.4.2. Inglise keelde tõlgitud teksti hindamine

Inglise keelde tõlkimise sisendina sai kasutatud Pool-korpust.

Tõlke sisend koosneb 2638 lausest, 15803 sõnast ning inimtõlge koosneb 23516 sõnast.

Google tõlge

Tõlge sooritatud Google SMT süsteemil (tehtud 1. veebruaril 2011)

```
Tulemused:

korreleerumine inimtõlkega
Sõna veamäär (kattuvuse %) 46.6727
Positsioonist sõltumatu sõna veamäär (kattuvuse %) 58.1264
BLEU meetrika  0.1764

puuduvad/liigsed unigrammid:
Liigsed 32.4622 %
Puuduvad 36.0752 %
Täpselt kattuvad 62.4328 %
keskmine puuduvate/liigsete n-grammide arv lausetes :
Liigseid:
unigramme 3.1566, bigrammme 7.6941, trigramme 7.8036, 4-gramme 7.4443
Puuduvaid:
unigramme 2.5504, bigrammme 6.4882, trigramme 6.5993, 4-gramme 6.2691

n-grammide järjekord:
Masintõlkes korrektseid n-gramme osakaal keskmiselt:
unigramme 0.6700, bigrammme 0.4028, trigramme 0.2998, 4-gramme 0.2513
Masintõlkes korrektsete n-grammide keskmine osakaal inimtõlkes:
unigramme 0.6292, bigrammme 0.3802, trigramme 0.2734, 4-gramme 0.2276

Vale sõna - sõna tõlkimisel tekkivad vead
Osaliselt kattuvad (sarnaste tüvedega) 1.4920 %
```

Hinnang:

Tõlge korreleerub päris hästi inimtõlkega, samuti on korrektsete sõnade ja bigrammide hulk päris suur. Sarnaste sõnade arv on päris väike, ligi 1.5 protsenti, seega on sõnalõpud enamasti hästi tõlgitud.

UT tõlge

Masintõlkesüsteem[3] on loodud Tartu Ülikooli Matemaatika-informaatikateaduskonna keeletehnoloogia teadusgrupi masintõlke töörühma poolt. Süsteemi eesmärgiks on demonstreerida grupisisest tööd laiemale avalikkusele ning saada kasutajatelt süsteemi väljundi kvaliteedi kohta tagasisidet.

```
Tulemused:

korreleerumine inimtõlkega
Sõna veamäär (kattuvuse %) 12.1745
Positsioonist sõltumatu sõna veamäär (kattuvuse %) 26.9093
BLEU meetrika 0.0815

puuduvad/liigsed unigrammid:
Liigsed 56.0810 %
Puuduvad 48.3575 %
Täpselt kattuvad 50.3726 %
keskmine puuduvate/liigsete n-grammide arv lausetes :
Liigseid:
unigramme 4.0227, bigrammme 9.4052, trigramme 9.2517, 4-gramme 8.6619
Puuduvaid:
unigramme 4.8472, bigrammme 10.5409, trigramme 10.8969, 4-gramme 10.2449

n-grammide järjekord:
Masintõlkes korrektseid n-gramme osakaal keskmiselt:
unigramme 0.4360, bigrammme 0.1906, trigramme 0.0941, 4-gramme 0.0853
Masintõlkes korrektsete n-grammide keskmine osakaal inimtõlkes:
unigramme 0.5086, bigrammme 0.2394, trigramme 0.1576, 4-gramme 0.1429

Vale sõna - sõna tõlkimisel tekkivad vead

Osaliselt kattuvad (sarnaste tüvedega) 1.2698 %
```

Hinnang:

Tõlge korreleerub eelmisest tõlkest kehvemini, samuti on korrektsete n-grammide hulk väiksem, samas on sarnaste sõnade arv natuke väiksem eelnevast tõlkest.

3.5. Tulemused

Tarkvara töötas nii nagu oodatud, selle tagamiseks tehti väikeste testandmete korral käsitsi kontroll. Lisaks sai lisatud BLEU meetrika arvutamine, mis näitab tõlgete korreleerumist. Negatiivseks aspektiks võib lugeda programmi keerukust ja ajanõudlikkust.

Python programmid töötavad tavaliselt aeglasemalt, kui vastavad Java programmid, kuid on see-eest tavaliselt 3-5 korda lühemad. Kuna iga lisatava lause kohta tehakse teatud hulk arvutusi, kasvab andmete lisamisel tehtava töö ja vaja mineva mälu hulk proportsionaalselt andmete mahu suurendamisega. Olulisemaks aeglustavaks teguriks pole mitte lausete hulk, vaid lausete, eriti just väga pikkade lausete esinemine tekstis, sest sama lause käiakse läbi mitu korda ning näiteks võimalike n-grammide arv kasvab eksponentsiaalselt lause pikkuse suhtes. Seega on näiteks ligikaudu sama hulga lausete töötlemiseks, kus ühel puhul oli 23000 ja teisel puhul 58000 sõna kulus ühe puhul ligi minut, teise puhul 6 minutit.

Programmi katsetati erinevate sisenditega, näiteks oli parem tulemus, kus sisendist oli eemaldatud kirjavahemärgid, või sisendiks antud eelnevalt morfoloogiliselt analüüsitud tekst.

Katsetest võib järeldada, et programmi katsetamiseks kasutatud tõlkesüsteemide tõlked korreleeruvad enam-vähem inimtõlkega, mida näitab ka BLEU skoor.

Kirjavahemärkide eemaldamisel paranes tõlke hinnet märgatavalt, seega on kirjavahemärkidega veel probleeme.

Seotud tööd

On palju erinevaid väljaandeid, mis tegelevad masintõlke erinevate automaatsete hindamismeetoditega. Osa neist pakuvad välja uusi meetmeid, teised parandusi ja laiendusi eksisteerivatele meetoditele. Näiteks 2006. aasta publikatsioonis pakuvad Maja Popovic jt. välja meetodi, mis kasutab morfoloogilist ja süntaktilist infot ning kombineerib seda automaatsete hinnangumeetoditega WER ja PER, et saada detailsemat infot tõlkevigade kohta. Selle meetodi puhul uuritakse kahte tüüpi vigu mis kaasnevas hispaania-inglise keelepaariga – süntaktilised erinevused kahe keele vahel, arvestades nimisõnu ja omadussõnu ning hispaania keele sõnade lõpud, arvestades põhiliselt tegusõnu, omadussõnu ja nimisõnu. Antud töö erineb eelnevalt kirjeldatud meetodist keelest sõltumatusega, ehk realiseeritud meetod ei sõltu keele morfoloogilisest ja süntaktilisest eripärast. Morfoloogilise info asendamiseks on töös sisse toodud üksiku sõna veamäär mis arvutab, kui mitu protsenti sõnast tuleks muuta, et saada ette antud sõna. Selle alusel võib pakkuda välja sõnadepaarid, mille tüved on sarnased ning erinevad ainult sõna lõpu poolest.

Kokkuvõte

Enamus kvaliteedi meetrikaid annab ainult mingi numbrilise väärtuse ette määratud vahemikust, andes masintõlkesüsteemide arendajate ainult üldise hinnangu süsteemi kvaliteedi suhtes, täpsustamata milliseid vigu analüüsi jooksul kohati. Käesoleva töö käigus prooviti lahendada seda mureküsimust - millist tüüpi vigu teeb süsteem tõlkimisel. Probleemi lahendamisel püüti säilitada keelest sõltumatust, mis omakorda seadis piiranguid programmi võimekusele teatud veatüüpe tuvastada. Näiteks ei saa keelespetsiifilisi andmeid omamata seada vastavusse inimtõlke ja masintõlke sõnu, mis ei ole sarnased. Töö lahenduse käigus prooviti läbi erinevaid lähenemisviise. Lõpuks jäi peale täielik keelest sõltumatuse nõue ehk loobuti morfoloogilise info kasutamisest ning arvutati keelest sõltumatult hinnangud erinevate vigade esinemise sagedustest. Programm suutis talle pandud eesmärke täita ning annab üpris adekvaatset statistikat süsteemi vigade kohta.

Masintõlge on väga oluline paljudele inimestele, selle abil saab kasutaja ligikaudese tõlke abil teksti sisust, mille jaoks oleks vaja muidu lingvisti. Kuna masintõlge on oluline, on ka selle arendamine oluline. Selleks on vaja süsteemi väljundit hinnata mingite kriteeriumite alusel. Alati on seda võimalik käsitsi teha, kuid see on aega ja ressurssi nõudev. Selleks on välja arendatud palju automaatsed hindamise meetrikad. Levinum neist on BLEU, mis hindab süsteemi tõlke ja näidistõlke korreleerumist. Veaanalüüs on siiski veel üsna uurimata ala ning loodetavasi sai töö autor anda oma panuse selle hüvanguks.

Translation quality evaluation

Master Thesis
Karol Toompalu
Abstract

The task of this thesis was to design a system that would help the machine translation system developers to identify what kind of errors their system makes while translating. To do that the analyzer generates detailed summary over the system output. The analyzer is able to identify missing and extra words/phrases, some differences in word inflections and differences in word/phrase order. It also calculates one of the most popular metrics BLEU value to help the developers to decide how well their system correlates with human translation.

The analyzer was tested on two languages English and Estonian. On each language two translation systems was chosen and their translation on the same input was compared to human translation. Examples of their output evaluation were presented in this thesis. It showed that overall quality of the systems was similar. Some differences occurred in the number of words with similar stems that occurred in translations. Error analysis is still a rather unexplored area and author hopes that he was able to put some contribution in the area.

Kirjandus

[1] Adam Lopez.: 2008, Statistical Machine Translation, ACM Computing Surveys 40(3), 1–49.

[2] David Vilar, Jia Xu, Luis Fernando D'Haro, Hermann Ney. 2006. Error analysis of machine translation output. In Proceedings of the 5th LREC, pages 697–702, Genoa, Italy.

[3] Eesti-inglise masintõlge, masintolge.ut.ee, (17.05.2011).

[4] Jesus Angel Gimenez Linares: 2009. „Empirical Machine Translation and its Evaluation" Ph.D. Thesis, Universitat Politecnica de Catalunya, Barcelona.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu: 2001, BLEU: a method for automatic evaluation of machine translation, in Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001), pp 311–318, Philadelphia, PA, USA.

[6] Mathias Madsen: 2009, "The Limits of Machine Translation", Master Thesis, University of Copenhagen.

[7] „Machine translation", Wikipedia, vaba entsüklopeedia, http://en.wikipedia.org/wiki/Machine_translation (15.05.2011).

[8] Michael Gilleland, „Levenshtein Distance, in Three Flavors", http://www.merriampark.com/ld.htm, (17.05.2011).

[9] NIST: 2002, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics", Technical report.

[10] Neil Coffey 2009: „Machine Translation - How it Works, What Users Expect, and What They Get", Ezine Articles, 8 Mai 2009.

[11] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer: 1993, The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics 19(2), 263–311.

[12] Ralf D. Brown: 2002, "Example-Based Machine Translation", tutorial at AMTA-2002.

[13] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 Mai 2006.

[14] Satanjeev Banerjee, Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL

Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65–72, Ann Arbor, Michigan, USA.

[15] Statistics Solutions ,“Correlation (Pearson, Kendall, Spearman) ”, http://www.statisticssolutions.com/resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman, (17.05.2011).

[16] Tiit Roosmaa, Kursuse „Masintõlge“ materjalid, http://www.cs.ut.ee/~roosmaa/MT98.html, (17.05.2011).

[17] Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, Omar Zaidan: 2009, Joshua: An Open Source Toolkit for Parsing-Based Machine Translation, in Proceedings of the Fourth Workshop on Statistical Machine Translation, pp 135–139, Athens, Greece.

Lisad

Lisa 1. Arhiiv

Käesoleval töö on üks lisa: arhiiv CD-plaadil, mis sisaldab probleemi lahendamiseks loodud koodi, katseandmed ning programmi kasutusjuhendit.